Assemblée annuelle de la SSC, juin 1999
Recueil de la Section des méthodes d'enquête

LE BIAIS DANS L'ÉCHANTILLON DU RECENSEMENT CANADIEN DE 1996

Sylvain Thivierge[1]

RÉSUMÉ

Dans le recensement de la population canadienne, l'information démographique de base est recueillie auprès de tous les ménages. De l'information additionnelle est recueillie auprès des ménages faisant partie d'un échantillon systématique sélectionné de façon indépendante dans chaque secteur de dénombrement au Canada. Tout chiffre de population dépendant de l'information recueillie dans l'échantillon seulement doit donc être estimé. Dans cet article, à l'aide de méthodes statistiques simples, nous montrons que l'échantillon du recensement est biaisé, en ce sens que certains types de ménages ou de personnes sont sur-représentés ou sous-représentés dans l'échantillon par rapport à la population, sans que cette différence échantillon/population puisse être expliquée seulement par la variabilité d'échantillonnage. La méthodologie utilisée pour détecter le biais y est décrite, et nous tentons aussi d'identifier certaines sources de ce biais. Puisque Statistique Canada publie des milliers d'estimations à partir des données du recensement, le biais dans l'échantillon du recensement est une préoccupation importante.

MOTS CLÉS : Recensement; estimation par calage; non-réponse totale; échantillonnage systématique.

ABSTRACT

In the Canadian Census of population, the basic demographic information is gathered on a 100% basis. Additional questions are asked on a sample basis to a one in five systematic sample of households, selected independently in each enumeration are in each enumeration area in Canada. Any population figure that depends on the information known for the sampled households only then has to be estimated. In this paper, using simple statistical techniques, we show that the Census sample is biased, in the sens that some household or person types are over-represented or under-represented in the sample compared to the population and that this misrepresentation cannot be explained by sampling variability only. The methodology used to detect the bias is described, and we also attempt to identify some sources of this bias. Since Statistics Canada publishes thousands of Census estimates, the bias in the Census sample is an important concern.

KEY WORDS: Census; calibration estimation; total non-response; systematic sampling.

1. INTRODUCTION

Dans le recensement de la population canadienne, chaque ménage doit fournir de l'information démographique de base sur toutes les personnes qui le composent. On demande ensuite aux ménages faisant partie d'un échantillon de fournir de l'information additionnelle. L'information de base est ensuite utilisée comme information auxiliaire pour estimer des caractéristiques de la population dépendant de l'information recueillie dans l'échantillon seulement. Après le recensement de 1996, comme après plusieurs recensements précédents, une importante étude de biais a été réalisée à Statistique Canada. L'objectif principal de cette étude était de déterminer si l'échantillon du recensement était biaisé, en ce sens que certains types de ménages ou de personnes étaient sur-représentés ou sous-représentés dans l'échantillon par rapport à la population, sans que ceci puisse être expliqué uniquement par le plan d'échantillonnage et la variabilité d'échantillonnage. Si un tel biais était présent, l'étude cherchait également à en identifier les causes.

Dans cet article, quelques résultats de l'étude de biais du recensement de 1996 sont présentés. L'accent sera mis sur la méthodologie utilisée puisque celle-ci diffère légèrement de celles utilisées dans les études de biais

[1] Sylvain Thivierge, Division des méthodes d'enquêtes-ménages, Statistique Canada, 16-F Immeuble R.H.Coats, Ottawa, Ontario, K1A 0T6, thivsyl@statcan.ca.

des recensements précédents. À la section 2, le plan d'échantillonnage et la méthode d'estimation utilisés dans le recensement de 1996 seront présentés. Plusieurs sources possibles de biais dans l'échantillon du recensement seront ensuite identifiées à la section 3. Finalement, la méthodologie utilisée pour détecter le biais dans l'échantillon sera décrite à la section 4, et quelques résultats de l'étude du biais seront présentés aux sections 5 et 6.

2. PLAN D'ÉCHANTILLONNAGE ET MÉTHODE D'ESTIMATION

Dans le recensement canadien de 1996, un échantillon systématique de ménages privés (avec pas de sondage égal à cinq) a été sélectionné dans 42 952 des 49 359 secteurs de dénombrement (SD) au Canada. Chaque ménage ne faisant pas partie de l'échantillon a reçu un questionnaire court, aussi appelé *questionnaire 2A*; ce document recueillait de l'information démographique de base sur les membres du ménage comme par exemple l'âge, le sexe et l'état matrimonial. Quant aux ménages échantillonnés, ils ont reçu un questionnaire long, aussi appelé *questionnaire 2B*. En plus de l'information de base recueillie par le questionnaire 2A, le questionnaire 2B recueillait de l'information supplémentaire portant, par exemple, sur le revenu, la langue, et le type de logement. L'information recueillie par les deux types de document est souvent appelée *information 2A*, alors que celle recueillie uniquement par le questionnaire long est appelée *information 2B*. Les ménages vivant en institution et les ménages vivant dans les SD où il n'y a pas d'échantillonnage (qui sont, dans la plupart des cas, des réserves autochtones ou des SD contenant exclusivement des ménages vivant en institution) ne sont pas considérés dans cet article.

L'information 2B n'étant connue que pour les ménages échantillonnés, tout chiffre de population en dépendant doit être estimé. La méthode d'estimation utilisée dans le recensement de 1996 était une méthode d'estimation par calage. Cette méthode consistait à ajuster le moins possible (par rapport à une certaine mesure de distance) les poids initiaux des ménages échantillonnés (qui sont définis comme étant l'inverse de la fraction d'échantillonnage pour tous les ménages échantillonnés d'un même SD) de manière à ce que les estimations de totaux, basées sur ces nouveaux poids, soient égales aux totaux correspondants pour plusieurs variables 2A appelées *contraintes*. Ces totaux furent calculés à partir de l'ensemble des ménages recensés. Des exemples de contraintes utilisées sont *le nombre d'hommes*, *le nombre de personnes*, *le nombre d'enfants âgés entre 0 et 4 ans* et *le nombre de personnes mariées*. Cet ajustement des poids fut réalisé indépendamment dans chacune des 5932 régions de pondération (RP), qui sont des regroupements de SD géographiquement contigus (sept en moyenne) contenant entre 1000 et 3000 ménages. Les poids obtenus à la suite de cet ajustement (que nous appellerons les poids finaux) furent utilisés pour produire toutes les estimations reliées à l'information 2B.

Selon le choix de la mesure de distance, on peut montrer que cette technique d'estimation est équivalente à une technique d'estimation par régression, dans laquelle les variables que nous avons appelées contraintes sont utilisées comme variables auxiliaires (voir Deville et Särndal, 1992, pour plus de détails). C'est précisément le cas de la mesure de distance choisie pour le recensement. Cette méthode est utilisée dans le recensement afin: 1) d'obtenir des estimateurs dont les variances sont plus petites que les estimateurs basés sur les poids initiaux, grâce à l'utilisation d'information auxiliaire; 2) de réduire l'impact de tout biais présent dans l'échantillon sur les estimateurs; 3) de réduire les différences entre les totaux de variables 2A et les estimations de ces totaux, qui apparaissent dans les tableaux présentant les estimations de totaux de variables 2A croisées avec des variables 2B. Le système de pondération du recensement de 1996 est décrit de façon détaillée dans Bankier, Houle et Luc (1997).

3. SOURCES DE BIAIS DANS LE RECENSEMENT

Le terme *échantillon biaisé* (ou *biais dans l'échantillon*) est peu habituel quoiqu'il soit souvent utilisé dans le contexte du recensement. Nous débutons cette section par expliquer ce que nous entendons exactement par *échantillon biaisé*.

L'échantillon originalement sélectionné dans le recensement est un échantillon systématique de ménages avec pas de sondage égal à cinq, stratifié par SD. En absence de toute source d'erreur, les poids initiaux des ménages échantillonnés, que nous avons définis à la section précédente, seraient égaux aux poids d'échantillonnage, c'est-à-dire cinq (pour simplifier les choses, nous supposerons que dans chaque SD, le nombre de ménages est un multiple de cinq). L'estimateur du total dans la population d'une variable d'intérêt, basé sur ces poids, serait alors sans biais (puisque cet estimateur serait simplement un estimateur d'Horvitz-Thompson). Dans ces conditions, l'estimateur d'un tel total, utilisant cette fois les poids calculés par le

système de pondération du recensement, serait approximativement sans biais. Toutefois, pour différentes raisons, le nombre de ménages 2B est différent du nombre de ménages originalement échantillonné dans la plupart des SD, et donc ces poids sont souvent différents de cinq. Puisque ce sont ces poids qui sont ajustés par le système de pondération plutôt que les poids égaux à cinq, nous dirons que l'échantillon du recensement est sans biais, si, pour n'importe quelle variable d'intérêt, l'estimateur du total dans la population de cette variable, basé sur les poids initiaux, est sans biais (sous l'hypothèse d'une répétition infinie du recensement et de toutes ses étapes sous les mêmes conditions qui prévalaient en 1996). Autrement, nous dirons que l'échantillon est biaisé.

Dans le recensement, plusieurs sources de biais potentielles peuvent être identifiées. Parmi les plus importantes, on retrouve:

1) les erreurs de couverture, qui surviennent lorsque certaines personnes ne sont pas recensées ou recensées plus d'une fois;

2) la conversion des questionnaires 2B en questionnaires 2A pour éliminer la non-réponse totale dans l'ensemble des ménages 2B (en d'autres termes, les ménages échantillonnés non-répondants sont exclus de l'échantillon);

3) le système d'imputation du recensement qui est utilisé pour éliminer la non-réponse partielle et corriger certaines réponses incohérentes;

4) les recenseurs qui ne distribuent pas toujours les bons questionnaires aux bons ménages;

5) les erreurs de réponse, qui surviennent lorsque certains ménages ne répondent pas correctement aux questions.

Notons que les points 2 et 4 expliquent le fait que les poids initiaux des ménages 2B ne sont pas toujours égaux à cinq. Ces poids initiaux peuvent être vus comme étant les poids d'échantillonnage, ajustés pour tenir compte de la différence entre la taille de l'échantillon visée et celle réellement obtenue.

4. MÉTHODE DE DÉTECTION DU BIAIS

Dans une région géographique donnée, supposons qu'il y ait G SD. Dans le $g^{\text{ième}}$ SD, dénotons par N_g le nombre de ménages et par n_g le nombre de ménages 2B, et considérons X_g, le total dans la population du SD d'une variable d'intérêt x (comme par exemple le nombre de femmes dans un ménage). Soit $\hat{X}_g$, l'estimateur de X_g qui utilise les poids initiaux pour pondérer les ménages 2B, c'est-à-dire l'estimateur défini par

$$\hat{X}_g = \frac{N_g}{n_g} \sum_{h=1}^{n_g} X_{gh} \ , \qquad (4.1)$$

où X_{gh} représente la valeur de la variable x pour le $h^{\text{ième}}$ ménage 2B . Dénotons maintenant par $\mathrm{Var}(\hat{X}_g)$ la variance de l'estimateur $\hat{X}_g$, calculée sous l'hypothèse que l'échantillon n'est pas biaisé. Finalement, posons

$$Z = \frac{\sum_{g=1}^{G} (\hat{X}_g - X_g)}{\sqrt{\sum_{g=1}^{G} \mathrm{Var}(\hat{X}_g)}} \ . \qquad (4.2)$$

Puisque les échantillons sont sélectionnés de façon indépendante dans chaque SD, la loi de la statistique Z ainsi définie devrait être très près d'une loi normale de moyenne 0 et de variance 1, si G n'est pas trop petit, ce qui sera le cas si la région géographique qui nous intéresse est une division de recensement, une province, ou encore le pays tout entier. Dans le cas où la variable d'intérêt x est une variable dépendant de l'information 2A, c'est-à-dire l'information recueillie auprès de tous les ménages, la statistique Z peut être calculée dans n'importe quelle région géographique, puisque la valeur d'une telle variable x est connue pour tous les ménages. La stratégie adoptée pour détecter le biais dans l'échantillon était donc de calculer la valeur de la statistique Z pour plusieurs variables 2A à différents niveaux géographiques pour ensuite évaluer leur plausibilité sous l'hypothèse qui stipule que la loi de Z est la $N(0, 1)$. Notons que la détection du biais pour les variables 2A suggère fortement que l'échantillon est aussi biaisé relativement aux variables 2B.

Évidemment, puisque les différentes sources d'erreur énumérées à la section précédente affectent autant les ménages 2A que les ménages 2B, les totaux X_g, que nous avons calculés à partir des ménages 2A et 2B, ne sont pas les "vrais" totaux. Nous les considérerons toutefois comme étant approximativement exacts. Tout comme pour les totaux X_g, les variances $\mathrm{Var}(\hat{X}_g)$ furent calculées en utilisant les données de l'ensemble

des ménages 2A et 2B, en supposant que le recensement n'était affecté par aucune source d'erreur. Il fallait toutefois tenir compte du fait que la taille de l'échantillon obtenue était souvent différente de celle visée. Pour ce faire, nous avons supposé que:

1) l'ensemble des ménages 2B ne faisant pas partie de l'échantillon original pouvait être considéré comme un sous-échantillon aléatoire simple sans remise sélectionné parmi les ménages non échantillonnés;

2) l'ensemble des ménages 2A faisant partie de l'échantillon original pouvait être considéré comme un sous-échantillon aléatoire simple sans remise sélectionné parmi les ménages échantillonnés.

Faire ces deux hypothèses revenait à supposer qu'il n'y avait pas de type de ménages ou de personnes qui était plus susceptible que d'autres de répondre ou de se voir assigner le mauvais type de questionnaire par un recenseur (et donc que les erreurs faites par les recenseurs et la non-réponse totale n'était pas des sources de biais dans l'échantillon).

Remarque. Dans l'étude du biais du recensement de 1991 (et dans les autres recensements précédents), la variance de $\hat{X}_g$ était calculée en supposant que dans chaque SD, l'échantillon originalement sélectionné était un échantillon aléatoire simple sans remise plutôt qu'un échantillon systématique. Toutefois, jamais le bien-fondé de cette hypothèse n'avait été étudié. En 1996, nous avons tenu compte de l'aspect systématique de l'échantillon, en identifiant dans chaque SD, les cinq échantillons qu'il était possible de sélectionner, et en calculant ensuite la variance de l'estimateur $\hat{X}_g$ à partir des cinq valeurs que celui-ci pouvait prendre. La variance ainsi calculée a ensuite été ajustée pour tenir compte des hypothèses 1 et 2 du paragraphe précédent.

Nous avons aussi calculé les variances des estimateurs basés sur les poids initiaux suivant les hypothèses de 1991 à plusieurs niveaux géographiques, afin de les comparer avec les variances calculées en tenant compte du plan d'échantillonnage systématique; il s'est avéré que les variances calculées en tenant compte du plan d'échantillonnage systématique étaient plus petites, en moyenne, et ce, à tous les niveaux géographiques. Toutefois, l'utilisation de l'un ou l'autre des ensembles de variances dans l'étude de biais n'en affectait pas les conclusions générales.

5. LE BIAIS AU NIVEAU DES DIVISIONS DE RECENSEMENT

Le biais dans l'échantillon du recensement a été étudié à plusieurs niveaux géographiques, à l'aide de la statistique Z définie à la section précédente, et ce, pour 32 variables démographiques reliées à l'âge, le sexe et l'état matrimonial des membres des ménages. Nous présentons ici les résultats au niveau des divisions de recensement (DR) pour quelques-unes de ces variables. Les DR, qui sont au nombre de 281 au Canada, sont des régions géographiques contenant, en moyenne, 153 SD et 38 000 ménages.

Pour chacune des 32 variables de l'étude, la statistique Z a d'abord été calculée dans les 281 DR. L'hypothèse stipulant que la loi de la statistique Z était la loi $N(0, 1)$ a ensuite été testée à l'aide du test de Kolmogorov au seuil de 5%. L'hypothèse fut rejetée pour 20 des 32 variables (63%). Le test de Kolmogorov fut appliqué une deuxième fois pour tester la normalité des statistiques Z, mais cette fois, sans spécifier la moyenne de la loi normale (tout en gardant la variance égale à 1); l'hypothèse ne fut rejetée pour aucune des 32 variables. Notons que ceci tend à confirmer que l'hypothèse de normalité de moyenne 0 et de variance 1 en l'absence de biais est raisonnable. Le biais dans l'échantillon semble donc affecter l'espérance de la statistique Z tout en préservant approximativement sa normalité (avec variance égale à 1). Le tableau suivant montre la moyenne et l'écart-type des statistiques Z ainsi que les seuils critiques ("p-values") des deux tests de Kolmogorov ($Prob > D$ et $Prob > D_2$) pour cinq variables.

Tableau 1: Moyenne et écart-type des statistiques Z et résultat du test de Kolmogorov pour quelques variables

Variable	Moyenne	Écart-Type	Prob > D	Prob > D_2
Nombre d'hommes	-0,07	1,07	0,68	0,99
Nombre de femmes	0,55	1,08	0	0,98
Nombre d'enfants (0 à 4 ans)	0,33	1,06	0	0,71
Nombre de personnes mariées	0,87	1,03	0	0,49
Nombre de personnes seules	-0,69	1,02	0	0,86

Ces résultats suggèrent que les femmes, les enfants de 0 à 4 ans et les personnes mariées étaient significativement sur-représentées dans l'échantillon du recensement, alors que les personnes vivant seules étaient significativement sous-représentées. Des exemples de d'autres types de personnes qui étaient significativement sur-représentés dans l'échantillon sont les enfants de 5 à 14 ans et les personnes âgées entre 45 et 54 ans, alors que les jeunes adultes (20 à 34 ans) et les personnes séparées (incluant divorcées et veuves) sont des exemples de types de personnes qui étaient significativement sous-représentés dans l'échantillon.

6. IMPACT DE LA NON-RÉPONSE TOTALE SUR LE BIAIS

Une des causes possibles de biais dans l'échantillon du recensement les plus facilement identifiables est la non-réponse. Comme dans la plupart des enquêtes, il y a deux types de non-réponse dans le recensement: la non-réponse totale et la non-réponse partielle. La première survient lorsqu'un ménage ne répond pas du tout à son questionnaire; la deuxième survient lorsque le ménage répond à au moins une question du questionnaire sans répondre à toutes. Nous analysons ici l'impact de la non-réponse totale et d'un certain type de non-réponse partielle sur le biais dans l'échantillon du recensement.

Lors du recensement de 1996, 86 183 ménages n'ont pas répondu du tout à leur questionnaire. Parmi ceux-ci, environ 40% étaient des ménages 2B. Tous ces ménages ont vu leurs questionnaires être transformés en questionnaires 2A et toute l'information 2A a été imputée (les ménages 2A non-répondants sont bien sûr demeurés des ménages 2A, et toute l'information 2A a été imputée). De plus, 3 358 ménages ayant reçu un questionnaire 2B ont répondu à au moins une question 2A, sans répondre à aucune question 2B; comme dans le cas de la non-réponse totale, ces ménages ont été transformés en ménage 2A, et l'information 2A manquante a été imputée.

Pour étudier l'impact de ces deux types de non-réponse, une analyse similaire à celle décrite à la section 4 fut faite. Nous en donnons quelques brefs résultats ici. Pour les 32 variables de l'étude, les statistiques Z ont été recalculées dans les 281 DR, mais cette fois, après avoir exclu tous les ménages qui n'avaient pas répondu du tout (autant dans l'ensemble des ménages 2B que dans l'ensemble des ménages 2A). De plus, tous les questionnaires 2B convertis en questionnaires 2A ont été reconvertis en questionnaires 2B. Ensuite, pour chacune des variables, la moyenne des 281 statistiques Z a été comparée avec la moyenne obtenue à la section 5.

Pour 22 variables sur 32, la moyenne des statistiques Z était considérablement plus proche de 0 que la moyenne correspondante obtenue sans exclure les non-répondants et sans reconvertir aucun questionnaire. Par exemple, la moyenne pour la variable *nombre de femmes* est passée de 0.55 à 0.43, et celle pour le nombre de personnes vivant seules est passée de -0.69 à -0.49. Cependant, pour 23 variables sur 32, la moyenne des statistiques Z était toujours significativement différente de 0, en se basant cette fois sur un test de Student à un seuil de 5%. À la suite de cette analyse, il fut conclu que les deux types de non-réponse dont nous avons parlés ici étaient une source importante de biais dans l'échantillon du recensement, sans pour autant en être la seule.

7. CONCLUSION

Les résultats de cette étude du biais dans le recensement canadien de 1996 ont clairement montré que l'échantillon du recensement (c'est-à-dire l'ensemble des ménages 2B) était biaisé, en ce sens que certains types de ménages ou de personnes y étaient sur-représentés ou sous-représentés par rapport à la population, sans que cette différence puisse être

expliquée par la variabilité d'échantillonnage seulement.

Dans cet article, nous avons identifié une source importante de biais dans l'échantillon: la non-réponse totale. D'autres sources potentielles de biais ont aussi été étudiées, comme par exemple le fait que les recenseurs n'attribuent pas toujours des questionnaires 2B à tous les ménages échantillonnés et uniquement aux ménages échantillonnés. Les contributions respectives au biais de ces autres sources potentielles se sont toutefois avérées trop petites pour être détectées. Une autre source de biais que l'on croit être importante et qui n'a toujours pas été étudiée et (mais qui le sera éventuellement) est le système d'imputation utilisé pour corriger la non-réponse partielle et les réponses incohérentes sur certains questionnaires, ainsi que pour éliminer la non-réponse totale sur les questionnaires 2A.

Même s'il est pratiquement impossible d'éliminer toutes les sources de biais dans l'échantillon, il est toutefois possible de réduire l'impact de quelques-unes d'elles sur les estimateurs du recensement. Par exemple, on peut montrer que la technique d'estimation par calage utilisée dans le recensement est susceptible de réduire l'impact de la non-réponse totale sur le biais des estimateurs, à condition que le modèle de régression sous-jacent soit approprié et que les totaux des variables 2A utilisées pour effectuer le calage soient les plus exacts possibles. Ces deux aspects seront étudiés dans le futur.

RÉFÉRENCES

Bankier, M., Houle, A.-M. and Luc, M. (1997). Calibration estimation in the 1991 and 1996 Canadian Censuses. *Procedings of the Survey Methods Section, American Statistical Association,* pp. 66-75.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, 376-382.


STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010305071